基于深度学习的视频行为识别技术综述 *

李 晨，何 明†，王 勇，罗 玲，韩 伟
(陆军工程大学 指挥控制工程学院，南京 210007)

摘 要： 行为识别(Action Recognition，AR)是计算机视觉领域的研究热点，在安防监控、自动驾驶、生产安全等领域具有广泛的应用前景。首先，对行为识别的内涵与外延进行了剖析，提出了面临的技术挑战问题；其次，从时间特征提取、高效率优化和长期特征捕获三个角度分析比较了行为识别的工作原理；再次，对近十年 43 种基准 AR 方法在 UCF101、HMDB51、Something-Something 和 Kinetics400 数据集上的性能表征进行比对，有助于针对不同应用场景选择适合的 AR 模型；最后，指明了行为识别领域的未来发展方向，研究成果可为视频特征提取和视觉内容理解提供理论参考和技术支撑。

关键词： 行为识别；深度学习；卷积神经网络；Transformer；RGB 视频
中图分类号： TP391.4　　doi: 10.19734/j.issn.1001-3695.2022.03.0077

Review of video action recognition technology based on deep learning

Li Chen, He Ming†, Wang Yong, Luo Ling, Han Wei
(*Command & Control Engineering College*, *Army Engineering University of PLA*, *Nanjing* 210007, *China*)

Abstract: Action recognition(AR) is a hot research area in computer vision field, and has an extensive application prospect for security monitoring, autopilot, production safety etc. Firstly, this paper analysed the connotation and denotation of AR and put forward the technical challenges; Secondly, the paper analysed and compared the working principles of AR from three aspects: time feature extraction, efficient optimization and long-term feature capture; Thirdly, in order to select suitable AR models for different application scenarios, this paper compared the performance characterization of 43 benchmark AR methods in recent ten years based on UCF101, HMDB51, Something-Something and Kinetics400 data sets. Finally, this paper pointed out the future development direction of AR field, and the research results can provide theoretical reference and technical support for video feature extraction and visual content understanding.

Key words: action recognition; deep learning; convolutional neural networks; Transformer; RGB video

0 引言

当前，视频数据成为信息的重要呈现形式，在各行业广泛应用。因此如何使计算机领会视频含义的视频理解技术逐渐成为研究热点。2017 年，计算机视觉与模式识别会议(CVPR)将视频理解划分定义为未修剪视频分类(Untrimmed Video Classification)、修剪动作识别(Trimmed Action Recognition)、时序行为提名(Temporal Action Proposal)、时序行为定位(Temporal Action Localization)、密集行为描述(Dense-Captioning Events)五项子任务[1]，本文行为识别(Action Recognition, AR)属于修剪动作识别范畴。针对 AR 中动作类别和任务的不同情况，理解 AR 内涵的侧重点也有所不同。在 Action 表示单人行为动作时(如跳跃、走路、攀爬等抽象事件)，动作粒度更细，分类模型需具备较强时间建模能力。在 Action 表示单人或多人行为活动时(如吃面包、踢足球等场景/对象事件)，识别模型可通过场景识别，时间推理能力要求较低。Recognition 有两种含义：a)classification，即对裁剪视频片段行为分类；b)detection，即给定未修剪视频，先定位行为始末时间，再进行分类。另外，输入数据又存在 RGB 视频图、骨骼图、深度图等多种形式。AR 研究领域存在以上概念的组合情况，但以分割 RGB 视频的行为分类为主，因此本文 AR 均指已修剪 RGB 视频的行为分类。

特征提取和分类是 AR 的核心问题，因视频是一组时间序列的图像帧，所以 AR 模型提取空间特征时还需考虑时间特征。

目前有两条特征提取思路：一是人工设计特征。此方法基于人对各特征的敏感程度，直接设计含有物理含义的特征提取器。其针对性较强，但存在忽视数据隐含信息和通用性差等问题。二是通过深度学习从数据中提取深度特征。此方法基于大脑皮层视觉理论设计模型结构，结合数据集和反向传播算法训练生成特征提取器。此种方式可应用于各类数据，但特征可解释性较差[2, 3]。学术领域对手工特征与深度特征两者谁更具优势尚未定论，考虑到目前分类任务以深度学习为主，因此本文围绕基于深度学习的 AR 模型论述。

目前基于深度学习的图像识别模型已走出实验室投入应用。AR 作为图像分类任务的时序扩展，在特征上多出时域信息需要提取，致其仍未实际部署。总结而言，AR 面临以下技术挑战问题：

a)视频数据集制作困难。识别精度提升需通过大量标注的数据集训练，但视频数据的标签注释、动作定位等工作非常费时，制约了视频数据集体量化和 AR 模型发展。

b)模型训练效率低。视频的数据量级较图像呈指数增长，导致 AR 模型拟合时空特征进行迭代优化的训练过程对硬件配置要求很高，需要大量时间。

c)类内高方差和类间低方差。AR 涵盖各类行为，同一类动作中各动作差别较大，而不同动作类又会呈现相似形式，

收稿日期： 2022-03-03；**修回日期：** 2022-04-19　**基金项目：** 江苏省重点研发计划资助项目；军内科研项目；军队重点课题资助项目
作者简介： 李晨(1994-)，男，山东巨野人，硕士研究生，主要研究方向为计算机视觉；何明(1978-)，男(通信作者)，新疆石河子人，教授，博导，博士，主要研究方向为智联网、无人指控等(paper_review@126.com)；王勇(1976-)，男，山东曲阜人，副教授，硕士，主要研究方向为指控装备；罗玲(1987-)，女，陕西安康人，讲师，硕士，主要研究方向为计算机视觉与模式识别；韩伟(1991-)，女，山东德州人，讲师，博士，主要研究方向为无人机集群协同控制.

这对 AR 特征提取器作出了更为精细的要求。

d)实时性较为不足。目前的 AR 模型为了追求高精度，轻量化工作较为滞后，另外基本在离线环境中仿真，视频都是预先修剪过的，很难对视频流在线识别行为。

国内外学者研究现状如下：刘勇等人[4]阐述了行为识别在智能家居中的应用流程；刘云等人[5]论述了基于深度学习的关节点行为识别方法；张晓平等人[6]从异常行为识别和异常行为检测两个角度对异常行为判别方法进行了分析；裴利沈等人[7]对传统方法和深度模型效果进行了对比分析。区别与以上研究工作，如图 1 所示，本文从时间特征提取、高效率优化、长期特征捕获三个角度对 AR 模型归纳，并总结了公共视频数据集以及主流和最新模型的性能对比。

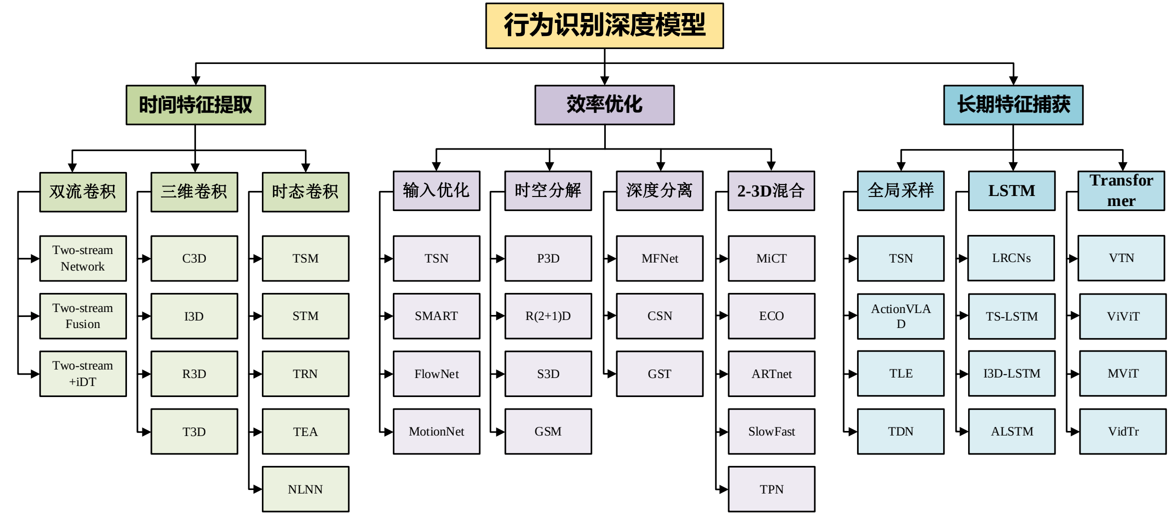


图 1　AR 模型分类总览图

Fig. 1　Overview of AR model classification

1　时空特征提取深度模型

AR 发展初期，以改进密集轨迹(iDT)[8]为代表的手工方法占据了主导，Hinton 等人[9]在 2015 年讨论了深度学习的原理和优势后，基于深度学习的 AR 工作逐渐展开。Karpathy 等人[10]基于卷神经网络(Convolutional Neural Network, CNN)，从堆叠视频帧中学习时空特征实现端到端动作分类，其评估了晚融合、早融合和慢融合等 2D CNN 连接方式，但识别精度远不及传统手工方法，表明此种简单的帧融合不能有效提取时域特征。

AR 较图像识别，不仅要关注空间特征，还要聚焦包括时域的时空特征从而理解运动信息。因此本节按照双流卷积、三维卷积以及时态建模三种策略，对 AR 的时空特征提取工作进行分析比较。

1.1　双流卷积模型

当人观察运动对象时，视网膜会流过连续图像，这些像素点的流动称为光流[11, 12]。光流通过表征图像变化携带运动信息，是提取时间特征的有效方法。Simonyan 等人[13]基于光流设计了双流网络(Two-stream Network)，两条 2D CNN 路径以视频帧和堆叠光流图为输入分别提取空间及时间特征。双流网络取得了与 iDT 比拟的识别效果，验证了光流对 AR 的有效性。Feichtenhofer 等人[14]基于双流网络探索了多种融合方式，并随着残差网络(ResNet)的推广，在文献[15, 16]中使用 ResNet 将双流连接，实现了时空信息的残差交互。在双流基础上，Wang 等人[17]基于 VGGNet-16 架构增加网络深度，并采用小学习速率、限制裁剪区域等方法缓解加深带来的过拟合问题。

鉴于双流网络的良好性能，文献[18]将双流网络的深度特征放置在 IDT 轨迹中心，构建了轨迹深度描述符(Trajectory-Pooled Deep-Convolutional Descriptors, TDD)。TDD 共享手工和深度特征，具有了更高区分度且能自动学习，此种融合方式成为 AR 刷新精度的有效方法。丁雪琴等人[19]对双流网络架构进行了改进，其将 BN-Inception 和 ResNet 引入，建立的时空异构双流网络验证了时空异构思想的有效性。

综上所述，双流网络使得深度学习方法在视频行为识别的地位获得极大提升，并逐渐发展成为 AR 的重要分支。

1.2　三维卷积模型

光流虽能提取时间特征，但易受光线变化影响，且对存储量和计算量要求较高，而且小位移特性也不易识别高速动作。图像识别中 2D 卷积取得了极好的效果，视频较图像多出时间维度，直接扩展 2D 卷积提取时空特征的工作得到开展。

Ji 等人[20]使用 3D 卷积核学习时空特征，证明了 3D 卷积在 AR 中的有效性。但他们未对 3D CNN 细致设计，识别精度不及双流网络和手工方法。后来 C3D[21]基于图像识别的 VGG-16 架构，使用 3×3×3 尺寸的 3D 卷积核取得了不错的识别效果。但 C3D 的精度较双流网络仍有差距，且参数量较大，在当时缺少大体量数据集的情况下训练周期长并易产生过拟合，另外存在的梯度消失/爆炸问题也限制了 C3D 的深度扩展。

鉴于 ResNet 能够缓解网络加深的退化问题，Tran 等人[22]设计了三维残差网络(3D Residual Networks, R3D)。R3D 将 ResNet 的 2D 卷积扩展为 3D，参数量较 C3D 降低了近 50%。后来 Hara 等人[23]又基于 R3D 进行深度扩展训练，对识别精度进一步提升。T3D[24]对 C3D 也进行了改进，但其使用的是 DenseNet 架构，参数量较 R3D 减少一半，但稠密连接会加大计算负荷。

研究初期的 3D CNN 一直未超越基于光流的双流网络，直到 2017 年 I3D[25]将困境打破。Carreira 等人[25]认为若把图像数据集中一张图片多次复制，就可生成一段“静态视频”训练 3D CNN。同理可将经过图像数据集预训练的 2D CNN 中的二维卷积核参数沿时间轴复制，便能得到初始化的 3D CNN，这为 AR 使用图像识别中的成熟架构提供了便利。他们将这种思想应用在双流网络的二维卷积路径，并首次使用 Kinetics 数据集进一步预训练，得到的膨胀三维卷积网络(I3D)比 C3D 网络更深，参数更少，成为了 AR 基准方法。

综上所述，3D CNN 逐渐超越了基于光流的双流网络，成为 AR 的另一重要分支。

1.3　时态卷积模型

双流网络和 3D CNN 的计算量普遍较高，不利于实时应用，且推理时间关系能力较弱。AR 模型需要随时间推移理解动作信息，因此一些研究聚焦于设计具有时态建模机制和低

计算量的时间模块，如图 2 所示。

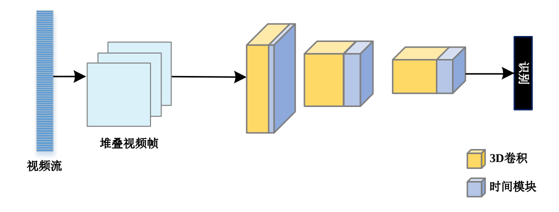


图 2　用于视频分类的时态卷积架构

Fig. 2　Temporal convolution architecture for video classification

时间关系网络(TRN)[26]在多个尺度上学习帧间时间关系，可即插即用到 CNN 架构中，但在输入帧数较多情况下，会导致模块太多造成训练困难。根据 3D 卷积可解耦为移步运算和乘法累加运算，时间位移模块(TSM)[27]将部分通道沿时间轴移位来提取帧间信息关系。TSM 模块可嵌入到各 2D CNN 识别模型中，在不增加计算情况下实现高效识别。TSM 的扩展工作 TIN[28]在通道维度上进行移位操作，并将移位操作的方向和开启设计为自动学习，精度上较 TSM 略微提升。TEI[29]模块通过分离通道相关和时间交互建模，TAM[30]使用动态时域卷积核自适应地聚合时域信息。时间激励聚合模块(TEA)[31]在 STM[32]基础上提出 ME 模块和 MAT 模块处理短程和长程特征。

罗会兰等人[33]设计了空间卷积注意力模块(SCA)和时间卷积注意力模块(TCA)。SCA 使用自注意力捕捉空间特征联系，用 1D 卷积提取时间特征。TCA 通过自注意力获取时间特征，用 2D 卷积学习空间特征。吴丽君[34]等人提出通道结合时间模块，通过调整池化层和卷积层的顺序,保留更多的有效通道信息和时间信息。

综上所述，时态卷积方法可将时空特征和运动特征整合到 2D CNN 中，不需要光流和三维卷积，具有时间建模能力同时消减了计算开销。

2　效率优化深度模型

1.3 节中，时态卷积模型具备时间建模的同时，彰显了较不错的效率优势。高效性是 AR 模型的重要指标，双流 CNN 中光流在存储和计算上是昂贵的，3D CNN 参数量和计算量较大，因此关于 AR 的效率优化任务得到开展。

2.1　输入数据优化

时域分段网络架构如图 3 所示。

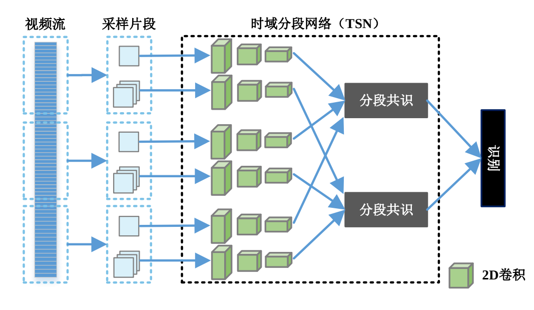


图 3　时域分段网络架构[35]

Fig. 3　Temporal segment networks architecture[35]

在输入帧方面，Wang 等人指出不是所有视频帧都包含有用信息，因此基于双流 CNN 提出均匀采样的时间段网络(TSN)[35]对视频帧均匀采样以提升效率。TSN 降低了信息冗余，以较低代价实现了端到端学习。关键帧挖掘框架[36]放弃随机策略，通过帧打分采样关键帧，但增益并不明显。文献[37, 38]认为帧对分类任务是有益的，他们将所有帧的前向输出进行聚合以提升效率。

针对光流不易计算问题，FlowNet[39]、FlowNet2.0[40]基于神经网络从图像中预测光流场，Piergiovanni 等人[41]基于 TV-L1 光流提出模拟光流的流卷积层，实现对光流迭代参数的端到端学习。隐式双流网络[42]将能够从视频帧中产生类光流的 MotionNet 与时间流 CNN 连接，缓解了光流计算开销。运动增强 RGB 流(MARS)[43]基于学习流思想，使用训练好的光流训练神经网络学习光流性能。Zhang 等人[44]通过运动边界的小位移解除对光流的依赖。

2.2　时空分解三维卷积

3D 卷积与 2D 卷积相比参数量和计算量大了很多，3D 卷积核的维度是 $F_C \times F_T \times F_H \times F_W$，其中 F_C 表示卷积核的通道数，$F_H \times F_W$ 表示卷积核的空间感受域，F_T 表示卷积核的时间感受域。在不考虑通道 F_C 的情况下，时空分解思想是将时空维度为 $F_T \times F_H \times F_W$ 的 3D 卷积核分解，近似为空间维度为 $1 \times F_H \times F_W$ 的 2D 卷积核和时间维度为 $F_T \times 1 \times 1$ 的 1D 卷积核的外积，如图 4 所示。

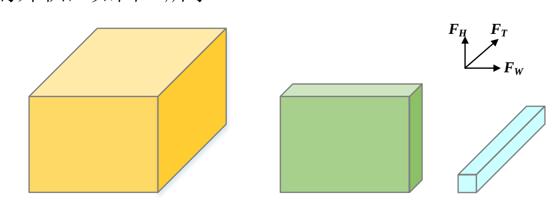


图 4　时空分解三维卷积

Fig. 4　Space-time decomposition of 3D convolution

基于此分解思想，P3D[45]使用 1×3×3 的 2D 卷积和 3×1×1 的 1D 卷积模拟 3×3×3 的 3D 卷积，P3D 参数量较 C3D 显著降低，且利用 2D CNN 初始化训练。Tran 等人[46]提出的 R(2+1)D，是和 P3D-A 类似的先 2D 卷积后 1D 卷积结构。但 R(2+1)D 利用效率优势增加通道数，准确率较 R3D 得到提升。S3D[47]采用 Top-heavy 方式简化特征量，优化了效率。近期，Sudhakaran 等人[48]提出 3D 时空分解的空间门控模块(GSM)，GSM 可通过时间自适应寻找特征并组合，几乎不需额外参数和计算。

时空分解具备效率优化的特点，但这种硬性的时空可分离方式会影响到 AR 模型的最优迭代，从而影响到 AR 的精细程度。

2.3　深度分离三维卷积

不同于时空分解卷积，深度分离卷积是将卷积核拆分为不同深度的卷积组。如图 5 所示，深度分离卷积是将维度为 $F_C \times F_T \times F_H \times F_W$ 的 3D 卷积核分解为两部分，一是 $1 \times F_T \times F_H \times F_W$ 的逐通道卷积核(Depthwise Convolution)，二是 $F_C \times 1 \times 1 \times 1$ 的逐点卷积核(Pointwise Convolution)，它将第一部分的特征在深度方向上加权组合生成特征，两个部分可在 Bottleneck 结构基础上共同作用优化模型效率。

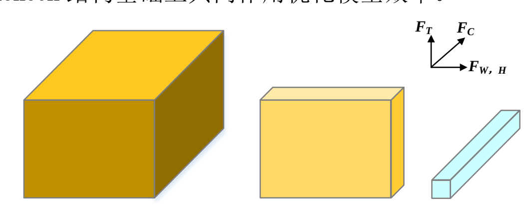


图 5　深度分离三维卷积

Fig. 5　Depth separation of 3D convolution

MFNet[49]基于 ResNet 和深度分离卷积思想，将 ResNet 模块切分为多纤维 ResNet 模块。实验证明 MFNet 计算量较 I3D 和 R(2+1)D 分别减少了 9 倍和 13 倍。通道分离卷积网络(CSN) [50]基于深度分离卷积，在 3D ResNet 模块上设计了三种 Bottleneck 结构，与 R(2+1)D 相比计算量减少了 2-3 倍。分组时空聚合(GST)[51]基于深度分离对 P3D 改进，对不同通道分别进行空间和时间操作以提升效率。

深度分离卷积能够减少参数量，但其中的逐通道卷积缺

少跨通道信息，导致缺乏空间关联，不利于 AR 模型的时空特征提取。

2.4 混合 2D 和 3D 卷积

鉴于卷积分解对识别效果的影响，联合 2D 和 3D 卷积的方法试图在保证精度的同时，进行效率优化。MiCT[52]在 3D 卷积后串联 2D CNN 延伸深度，另外并行 2D CNN 避免深度增加造成的梯度消失和训练误差，有效控制了 3D CNN 复杂性。相反，高效卷积网络(ECO)[53]通过 2D CNN 得到特征图后，再连接 3D CNN 实现分类。ECO 支持快速处理，能在 1 秒内进行 230 段视频的动作分类。

ARTnet[54]基于双流思想，双流分别配置 2D 和 3D 卷积提取空间和时间特征。SlowFast[55]网络类似于 ARTnet 的双流路径，但 SlowFast 设计了慢-快路径。如图 6 所示，慢路径聚焦空间特征，输入上使用低帧采样和较大的通道数，约占 80%计算量。快路径聚焦时间特征，输入上使用高帧采样和较小的通道数，约占 20%模型计算量。但行为的节奏多样，SlowFast 需要设置不同帧率，且事先定义不同帧率并不实际。针对此问题，时间金字塔网络(TPN)[56]在使用一个帧率情况下，提取不同层次的金字塔式特征图，表征各速率特征；BQN[57]将快慢信息自动分开，通用性更强。刘钊等人[58]为了降低 3D CNN 的参数量提出了时域零填充卷积网络，其先以时域不填充的方式使用 3D 卷积提取时空信息，然后利用网络重组结构将 3D 卷积变为 2D 来进一步提取特征。

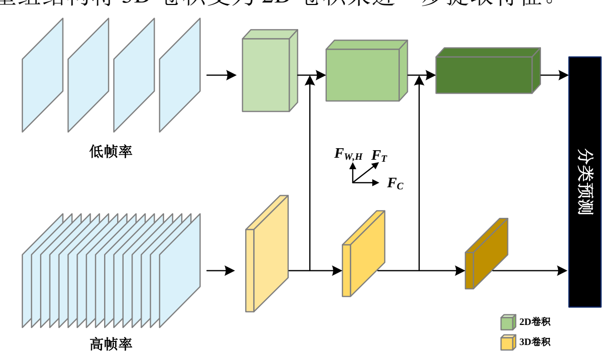


图 6 用于行为识别的 SlowFast 网络[55]

Fig. 6 Slowfast network for action recognition[55]

综上所述，AR 效率优化都会在深度、空间、时间、通道、采样等指标上扩展或压缩，但人工设置对精度和效率的平衡是次优的。最近 X3D[59]基于各指标自动逐步扩展和评估反馈，在实现良好精度的同时极大提升了运行效率。MoViNet[60]不同于 X3D 的定义扩展，使用神经网络搜索生成高效和多样化的 3D CNN，并取得了极好的效率-精度平衡。

3 长期特征捕获深度模型

前文模型提取的是短期动作特征，对于起止间隔较长的动作(如跳高和跳远)识别效果较差。长期时间卷积(LTC)[61]堆叠更多视频帧增强长期特征性能，FOF[39]、FCF[40]叠加多个表示流层捕获更长时间特征。但这些方法算量较大，并且长间隔帧间关系易丢失，因此研究者针对如何捕获长期行为特征的问题进行了研究。

3.1 全局均匀采样

第 2 节中 TSN[35]使用稀疏采样策略固定了计算量，但是同时此策略还得到了全局采样帧，实现了长期特征提取。因此稀疏采样策略，被 AR 模型的数据预处理阶段广泛采用。

但 TSN 仅将采样帧预测得分平均，不能弥补虚假标签损失。Lan 等人[62]将特征聚合成全局特征后，在相同训练数据上训练出映射函数，从而将全局特征映射到全局标签。ActionVLAD[63]将双流时空特征做池化聚合，实现了全局特征的整合。Diba 等人[64]将采样特征融合进行时间线性编码(TLE)，捕获长时间动态过程。

Wang 等人[65]基于 TSN 提出了时序差分网络(TDN)，TDN 设计了基于不同特征的通道注意力增强方法，实现对段间运动变化信息的增强。

3.2 长短时记忆网络

长短时记忆(LSTM)[66]在表征语言序列上效果显著，具备较强的长期特征捕获能力。视频具有和语言类似的时间上下文关系，因此 Srivastava 等人[67]认为 LSTM 是促进 AR 模型学习长序列关系的有效途径。

如图 7 所示，Ng 等人[68]先使用 2D CNN 提取空间特征，再输入 LSTM 进行融合实现时序特征提取。在此基础，长期递归卷积网络(LRCNs)[69]进行了端到端训练的优化工作。TS-LSTM[70]将特征矩阵划分为若干时间段，分别平均或最大池化汇集，按顺序输入 LSTM 层。I3D-LSTM[71]基于 I3D，对 3D CNN 和 LSTM 的结合工作作出尝试。Li 等人[72]将 LSTM 的权重点积改成卷积运算，证明 Conv-LSTM 较 LSTM 更有利于注意力机制发挥。

LSTM 一定程度增强了 CNN 的长时表征能力，但 LSTM 本身训练比较困难，再加时序先后顺序的严格迭代比较影响训练效率。

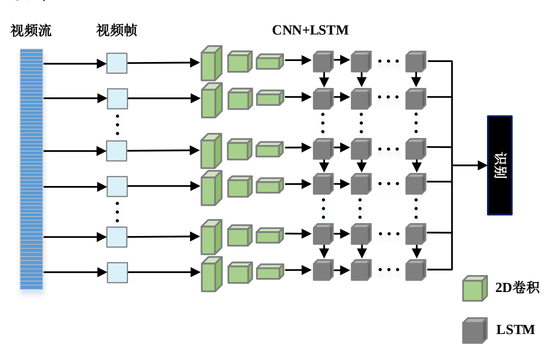


图 7 基于长短时记忆网络的行为识别模型[68]

Fig. 7 Action recognition model based on LSTM[68]

3.3 Transformer

CNN 和 LSTM 只有通过重复堆叠才能捕获长期依赖关系，但同时特征也会随距离增加逐渐衰减，并且运算开销也较大。2017 年谷歌在自然语言处理领域提出 Transformer[73]，Transformer 不管序列间距离有多远，其多头自注意力机制都能直接关注到任意序列间的全局信息，在运算上具备很强的并行性。Wang 等人[74]基于自注意力机制提出了非局部神经网络(NLNN)，NLNN 能够计算任意两个时空位置间的关系，从而快速捕获长期特征。Neimark 等人[75]提出了基于 CNN+Transformer 的 AR 模型 VTN，其利用 2D CNN 提取特征后，再通过 Transformer 结构关注长期信息。UniFormer[76]基于时空自注意力，分别在浅层和深层 CNN 学习局部和全局标签相似性，来解决时空冗余和依赖关系，在计算和准确性之间取得了更好的平衡。

ViViT[77]基于 ViT[78]完全摈弃 CNN，使用纯 Transformer 进行 AR 任务。如图 8 所示，ViViT 将视频构建为一组时空标签和时空位置编码后，作为 Transformer 的输入进行分类任务。MViT[79]基于 ViT 创建多尺度特征金字塔，首先在高分辨率下建模低层次视觉信息，后来在低分辨率下建模复杂高维特征。Li 等人[80]对 MViT 作出改进，分解了相对位置嵌入和残余池连接。由于视频帧之间存在较大的局部冗余和复杂的全局依赖性，VidTr[81]和 STAM-32[82]受卷积分解启发，基于 ViT 提出可分离注意分别执行空间注意和时间注意，减少了编码的计算消耗。

同一组视频帧若在时间排序不同，可能会表征不同动作，例如走路可能会变成跑步。然而传统的注意力机制不包含相

关性的方向信息，因此 DirecFormer[83]基于余弦相似度将 Transformer 中的注意力机制改造为定向时间注意和定向空间注意力，以正确的顺序理解人类行为。

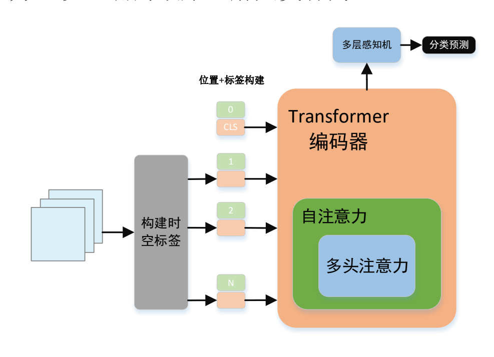


图 8　基于 Transformer 的行为识别模型[77]

Fig. 8　Action recognition model based on Transformer[77]

BEVT[84]开展了 BERT 预训练工作用于 AR 任务，它采用解耦设计首先对图像数据进行掩码图像建模，然后通过权重共享对图像和视频数据进行联合掩码图像建模和掩码视频建模。BEVT 简化了 AR Transformer 的学习，并且保留了从图像中学习的空间知识。

因 Transformer 在各类数据上的通用性，基于 Transformer 的多模态 AR 研究得到发展。Alfasly 等人[85]采用 BERT 构建语义音频视频标签字典(SAVLD)。SAVLD 将视频标签映射到其最相关的音频标签，然后与预训练的音频多标签模型在训练阶段共同估计视听模态的相关性。Zellers 等人[86]设计了适用于所有模式和时间步长的联合编码器(Transformer)，通过向联合编码器提供视频帧以及单词或音频的序列级表示，以预测数据内容。

2021 年开始，基于 Transformer 的 AR 模型持续刷新各基准数据集的精度榜单，具备极好的长期特征捕获能力。但是 Transformer 模型缺乏归纳偏置能力，不具备 CNN 的平移不变性和局部性，因此在数据不足时不能很好的泛化到 AR 任务上。

4　深度模型评估

本节在 4.1 节介绍了公共视频数据集。在 4.2 节和 4.3 节，基于 UCF101、HMDB51、Kinetics400 和 Something-Something 数据集，对各 AR 模型的识别精度和运行效率进行了分析比对。

4.1　视频数据集

高效准确的模型设计是 AR 的核心工作，而视频数据也同样重要。视频数据集应符合类别均衡、数据充足、标记正确、任务相关等特点。Duan 等人[87]利用 Web 数据训练视频识别模型，克服了数据格式之间的障碍，Zhang 等人[88]将 Transformer 在不同视频数据集联合训练学习到更好的动作表示，Ryoo[89]在视觉数据中学习挖掘数据集标签制作数据集，通过此数据集训练的 TokenLearner 获得了极好的识别效果。以此可见数据集对 AR 模型的重要作用，所以本节表 1 对 AR 任务中的 16 种公共数据集进行介绍。

HMDB51[90]来自公共数据库，共 6,849 个剪辑视频，分 51 个动作类，类别包括面部动作、一般身体动作、物体交互动作、人类互动动作。

UCF101[91]从 YouTube 收集，共 13,320 个剪辑视频，视频分为 25 组，每组包含 4-7 个动作类，动作类别包括人物交互、单人动作、人人交互、演奏乐器、运动。

Kinetics 是目前 AR 主要采用的数据集。第 1 代 Kinetics-400[92]数据集来自于 YouTube 视频，包含 400 个动作类，共 306,245 个视频；第 2 代 Kinetics-600[93]扩展到 600 个动作类，共 482,622 个视频；第 3 代 Kinetics-700[94]将动作类扩展到 700，共 650,317 个视频；Kinetics-700-2020[95]对 700 类扩展到至少 700 个剪辑视频。

Something-Something[96](Sth-Sth)数据集带有大量动作标签，更强调动作本身，包含人对日常对象执行的基本动作，动作类有 174 个。V1 版本有 108,499 个视频，V2 版本中有 220,847 个视频，持续时间 2 秒到 6 秒。

表 1　AR 数据集对比

Tab. 1　Comparison of AR data sets

年份	数据集	动作类别	剪辑数	应用
2009	Hollywood$_2$[97]	12	3,669	电影中动作
2011	HMDB51[87]	51	6,849	身体交互动作
2012	UCF101[88]	101	13,320	交互动作
2014	Sports-1M[98]	487	1,000,000	体育视频
2015	ActivityNet[99]	200	19,994	日常动作
2020	Kinetics[92]	700	650,000	YouTube 视频
2019	Charades[100]	157	9,848	日常行为
2019	Moments in Time[101]	313	1,020,000	动词动作标签
2017	Sth-Sth[93]	174	108,499	日常基本动作
2020	TITAN[102]	50	700	车辆、行人动作
2019	20BN-JESTER[103]	27	148,092	人类手势
2019	MMAt[104]	37	36,000	多模态动作
2020	RareAct[105]	122	7607	异常交互动作
2021	TinyVIRAT[106]	26	12,829	监控视频
2021	UAV-Human[107]	119	67,428	无人机视角动作
2021	Action Genome[108]	75	1750	组合动作

Action Genome[108]是一个具有多种模式和视点的多视图动作数据集，辅以分层活动和原子动作标签以及密集的场景组合标签，具有高级活动和低级动作的定义。

4.2　精度评估

本节对各 AR 模型的参数量、计算量、训练次数、数据预处理、软硬件配置等方面不做考虑，聚焦模型的单标签数据集精确度(accuracy)，为 4.3 节效率评估的模型分析提供参考。表 2 引用原文中各方法在 UCF101 和 HMD51 数据集的精确度，按照时间、技术原理横纵向排列比对，并标明了光流、架构使用和预训练情况。

通过表 2 可知：Slow fusion CNN[10]在 AR 中的早期尝试效果并不理想，这是因为 2D CNN 缺乏时间特征提取能力；基于光流的 Two stream CNN[13]表现出很好的识别效果，说明光流时间特征对 AR 的积极作用；同时，基于 3D CNN 的 C3D[21~23, 25, 61]等模型证明了 3D 卷积核提取时空特征的有效性；另外，TSM[27, 31, 32, 57]等模块对时态单独建模，彰显出时间模块对于 CNN 时间表征的强大性能；反之，LSTM 对 CNN 的 AR 精度增益非常有限[68, 70, 71]；最后，摈弃 CNN 使用纯 Transformer 的 AR 模型[81, 82]，取得了与其他技术代表模型 SMART[38]、Two-Stream I3D[25]、BQN[57]和 I3D-LSTM[71]比肩的识别精度。因此，目前 AR 实现可着重关注基于 Transformer 或者时间模块的特征提取技术。

对于表 2 各部分横向对比：双流部分中，TDD+IDT[18]较 Two-stream[11]略微增益，Fusion two-stream[14]和 ST-ResNet[15]表明，双流间融合和结合 ResNet 增加深度是提高双流精度的合适方法；3D 和时间模块部分中，R3D[22]、ResNeXt[23]、Two-Stream I3D[25]和 TSM[27]、TEA[31]、BQN[57]等 AR 模型，都使用 ResNet 或 BN-Inception 架构加深卷积层以提高精度；采样和时空分解部分表明，稀疏采样或横向压缩模型体量，进而纵向延伸网络深度的方式是提升精度的有效选择[35, 38, 46, 47, 64]。

结合表 2 中预训练指标，观察到基于 2D CNN 的 Two-stream[13]、TSN[35]等双流和时间模块部分均使用 ImageNet 数据集预训练。3D 部分中 I3D[25]提出膨胀思想后也能够使用

ImageNet 数据集预训练，并且首次使用 Kinetics 视频数据集预训练。I3D 取得了极好的识别效果，后来的各 AR 方法均采用了与 I3D 类似的预训练方式。这表明数据集对于 AR 精度提升的重要性，例如 Omni[87]的大数据联合统计训练的新型训练方法展现出极好的识别效果。

表 2　各 AR 模型在 UCF101 与 HMD51 数据集上识别精确度比较

Tab. 2　Comparison of identification accuracy of each AR model on UCF101 and HMD51 data sets

年份	模型	光流	预训练	主干架构	UCF101	HMD51
2014	Slow fusion CNN	-	ImageNet	AlexNet	65.4	-
2014	Two-stream	+	ImageNet	VGG-M-2048	88.0	59.4
2015	TDD+iDT	+	ImageNet	VGG-M-2048	91.5	65.9
2016	Fusion two-stream	+	ImageNet	VGGNet-16	92.5	65.4
2016	ST-ResNet+iDT	+	ImageNet	ResNet50	94.6	70.3
2014	C3D	-	Sports-1M	VGG-11	85.2	56.8
2016	LTC	-	Sports-1M	VGG-11	91.7	64.8
2017	R3D	-	Sports-1M	ResNet-18	85.8	54.9
2017	Two-Stream I3D	-	ImageNet+Kinetics	BN-Inception	98	80.9
2018	ResNeXt	-	Kinetics	ResNet-101	94.5	74.5
2019	TSM	-	ImageNet	ResNet-50	95.9	73.5
2019	STM	-	ImageNet+Kinetics	ResNet-50	96.2	72.2
2020	TEA	-	ImageNet+Kinetics	ResNet-50	96.9	73.3
2021	BQN		ImageNet+Kinetics	ResNet-50+TSM	97.6	77.6
2016	TSN	+	ImageNet	BN-Inception	94.2	69.4
2017	TLE	+	ImageNet	BN-Inception	95.6	71.1
2020	SMART	-	Kinetics	ResNet-152	98.64	84.36
2017	S3D	-	ImageNet+Kinetics	BN-Inception	96.8	75.9
2018	R(2+1)D	-	Kinetics	ResNet-34	96.8	74.5
2015	Two-stream+LSTM	+	ImageNet	VGGNet-16	88.6	-
2019	TS-LSTM	+	ImageNet	ResNet-101	94.1	69
2019	I3D-LSTM	-	Kinetics	BN-Inception	95.1	-
2021	VidTr-Lr	-	Kinetics	ViT-B	96.7	74.4
2021	STAM-32	-	ImageNet+Kinetics	ViT-B	97	-
2021	Omni	-	Kinetics+OmniSource	ResNet-101+I3D	98.6	83.3

UCF101、HMD51 和 Kinetics 数据集中，视频帧的动作与场景具有较强相关性。因此 AR 模型在此类数据集上的高精度并不能完全验证时间建模能力，所以需要在侧重动作特征的 Sth-Sth 数据集上对模型评估。不考虑参数计算量、训练次数、数据预处理、软硬件配置等因素，表 3 引用原文中各方法在 Kinetics400 和 Sth-Sth 数据集的精准度，按照时间、技术分区横纵向排列，并标明了架构使用情况。

通过表 3 可知：首先，TSN[35]和 I3D[25]在 Kinetics400 上性能相近，但在 Sth-Sth 上 TSN 与 I3D 存在较大差距。这说明 TSN 的稀疏采样策略丢失了大量运动信息。其次，相同 ResNet-50 架构下，第一部分的 TSN 和 I3D 较第二部分的 TSM[27]、STM[32]、TEA[31]在 Kinetics400 上识别精度并不占优势，在 Sth-Sth 上差距更被拉大。另外，相同 ResNet-101 架构下，CSN[50]和 SlowFast[55]较 PAN[44]、TDN[65]、BQN[57]在 Kinetics400 上识别精度略占优势，但在 Sth-Sth 上又再次被拉开距离。说明与 3D 卷积核和光流相比，在 CNN 上单独设计时间模块确实更能有效提取运动特征。最后，2021 年兴起的基于 Transformer 的 AR 模型在 Kinetics400 和 Sth-Sth 上持续刷新精度榜，直接超越了发展多年的各 CNN 模型。

综上所述，若 AR 应用于场景相关任务，需要关注 Transformer 技术、时间模块设计、横向压缩体量及残差连接和大体量数据集预训练等方面。若应用于识别场景弱化的动作相关任务，不要选择间隔过大的采样策略，另外可聚焦时间模块或 Transformer 技术进行模型设计。

表 3　各 AR 模型在 Kinetics-400、Sth-Sth V1 数据集上识别精确度比较

Tab. 3　The identification accuracy of AR models was compared on Kinetics-400 and Sth-Sth V1 datasets

年份	模型	主干架构	Kinetics-400		Sth-Sth V1		Sth-Sth V2	
			Top1	Top5	Top1	Top5	Top1	Top5
2016	TSN[35]	ResNet-50	73.9	91.1	19.7	46.6	30.0	60.5
2017	I3D[25]	ResNet-50	72.1	90.3	41.6	72.2	43.8	73.2
2017	S3D-G[47]	BN-Inception	77.2	93.0	48.2	78.7	69.4	89.1
2019	CSN[50]	ResNet-101	82.6	-	53.3	-	60.5	-
2019	GSM[48]	BN-Inception	77.5	-	55.16	-	62.7	-
2019	SlowFast+NL[55]	ResNet-101	79.8	93.9	-	-	61.7	-
2017	TRN[26]	BN-Inception	72.5	-	42.01	-	55.52	83.06
2017	Non-local[74]	ResNet-50+I3D	77.7	93.3	44.0	76	-	-
2018	TSM[27]	ResNet-50	74.7	-	50.7	-	66.6	91.3
2019	STM[32]	ResNet-50	73.7	-	50.7	80.4	64,2	89.8
2020	TEA[31]	ResNet-50	76.1	92.5	52.3	81.9	64.5	89.8-
2020	PAN[44]	ResNet101	77.3	-	55.3	82.8	66.5	90.6
2020	TDN[65]	ResNet-101	79.4	94.4	56.8	84.1	69.6	92.2
2021	BQN[57]	ResNet-101	77.3	93.2	57.1	84.2	-	-
2021	UniFormer-B[76]	I3D+Transformer	82.9	94.5	60.9	87.3	71.2	92.8
2021	MViT-L[80]	ViT-B	86.1	97.0	-	-	73.3	94.1
2021	MaskFeat	ViT-B+MViT-L	87.0	97.4	-	-	75.0	95.0
2021	ViViT[77]	ViT-B	84.8	95.8	-	-	65.4	89.8
2021	CoVeR[88]	ViT-B	87.2	97.5	-	-	70.9	92.5

4.3　效率评估

4.2 节基于识别精度对各模型的时间建模能力进行了对比分析，但评估 AR 模型需注重效率以便应用。不考虑训练次数、数据预处理、软硬件配置等因素，本节表 4 引用各 AR 模型原文中的预训练情况、架构使用、输入帧数、参数量(参数量代表占用显存量)、GFLOPS(代表执行时间的长短，要求在于 GPU 的运算能力)和精确度指标，对各基准模型进行效率评估。

通过表 4 可知：首先，TSN[35]表明 ResNet 较 BN-Inception 识别精度更高的情况下 FLOPS 更小，识别精度更高(ResNet 的 8 帧强于 Inception 的 25 帧)。因此目前的 AR 模型普遍采用 ResNet 作为基础架构，但 ResNet 对参数量提出了更高要求。其次，在 3D CNN 的效率优化工作中，时空分解的 S3D-G[47]较 I3D 执行次数得到显著降低，并且精度得到略微提升；ARTNet[54]和 MF-Net[49]分解工作只提升了模型效率而损失了精度；SlowFast[55]为了保证精度，模型参数量和计算量也并不可观；最近 X3D[59]使用逐步扩展模型和 MoViNet[60]网络架构搜索方法取得了极好的效率和精度平衡。另外，以 TSM[27]为代表的时间卷积模型体量与分解 3D 卷积模型相当，并在 Kinetics400 保持了稳定的识别精度。如 TDN[65]在不增加 TSN 体量的前提下，识别精度得到极大提升，这证明了插入时间模块方式的高效率和强大的时间建模能力。最后，基于 Transformer 的 AR 模型[77, 79, 80, 89]在保证不增加计算量的情况下，将识别精度突破到了 80%以上，识别性能超过了基于 CNN 的大部分模型。

综上所述，若 AR 应用于在线任务，需要关注卷积分解、时间模块设计、Transformer 等方面。但是基于 Ttransformer 的 AR 模型需要大量数据训练才能发挥效果，一些数据量较小的应用很难让 Ttransformer 发挥作用。利用迁移学习是解决此问题的合适途径。

5　结束语

本文从时间特征提取、高效率优化、长期特征捕获三个角度对 AR 模型分析，并在介绍公共视频数据集后对比了各

基准模型的精度和效率性能。虽当前 AR 模型在各公共数据集上表现良好，但距离实际应用仍有差距。以下是本文对 AR 领域未来发展方向的参考性见解。

a)小样本学习。训练 AR 模型需要大量标签视频，而视频标签注释的成本巨大，这造成基于监督学习的 AR 模型难以实际应用。另外，因环境背景的不同，也会影响不同环境中训练的 AR 模型。因此，涉及跨数据集的跨域学习、迁移学习和无监督学习等小样本学习有利于缓解标注成本，同时提高通用性。 如充分聚合时空上下文利用有限样本[109]、将图像数据集转换为视频模型预训练数据源[110]、使用未标记视频进行预训练[111]等方法。

b)视频语义理解。目前的 AR 方法是直接提取单动作特征，而实际的人类行为是一种复杂活动，如正在发生什么行为、行为何时发生、谁在执行行为以及行为发生在哪里。因此当识别复合行为时，不仅要利用分类模型，还需注重视频内容的语义理解。从视频数据中生成基本语义进而理解复杂语义，是弥补低级与高级行为间含义差的有效途径。

c)细粒度行为识别。细粒度行为识别需关注细微的时空语义差异，例如一个人是在缓慢走路还是快速走路。了解行为的细节执行方式，设计出能表示行为是如何发生的 AR 特征提取器，以更好区分细粒度行为类别，是值得研究的方向。

d)多模态行为识别。人类通过处理多种模态信息感知环境，如音频、触觉、视觉和骨架等，模态间形式不同且相互补充。AR 可在关注视觉信息的基础上，基于多模态数据研究如何在训练时利用多模态数据的互补性，以便学习出更好的 AR 特征提取器。

e)多视图行为识别。目前 AR 主要针对视频的单视图，但在实际应用场景中，摄像头被放置在不同方向，所获取的信息角度也是不同的。这种多视图数据给 AR 带来了挑战，同时也获得了机遇。将多视图数据进行三维重建，构建全方位的三维信息，进一步设计基于三维视频数据的特征提取器是未来值得探索的方向。

f)高效的模型开发。在实际应用中，AR 技术需满足处理速度快、计算成本低、存储空间小等要求，前文总结的效率优化方法大都是人工优化。通过神经架构搜索生成高效多样架构，从而高效集成，是优化 AR 效率的未来方向。

表 4　各 AR 模型在 Kinetics-400 数据集上的效率评估，view(时间剪辑数×空间剪辑数)；计算量 FLOPs，FLOP 指浮点运算次数，s 是指秒，即每秒浮点运算次数，考量一个网络模型的计算量的标准；参数量是指网络模型中需要训练的参数总数

Tab. 4　Efficiency evaluation of AR models on Kinetics-400, ‘view’(number of temporal clips×number of spatial clips); ‘flops’ refers to the number of floating point operations. ‘s’ refers to the number of floating point operations per second. ‘Parameter’ refers to the total number of parameters that need to be trained in the network model

年份	模型	预训练	主干架构	帧数×view	参数量/M	GFLOPS×view	Kinetics400	
							Top1	Top5
2016	TSN	ImageNet	BN-Inception	25×10×1	10.7	53×10×1	69.1	88.7
2016	TSN	ImageNet	ResNet-50	8×10×1	24.3	33×10×1	70.6	89.2
2017	I3D	ImageNet	BN-Inception	64×N/A×N/A	12	108×N/A	72.1	90.3
2017	S3D-G	ImageNet	BN-Inception	64×10×3	11.56	71.38×10×3	74.7	93.4
2017	ARTNet	ImageNet	ResNet18	16×25×10	35.2	23.7×25×10	70.7	89.3
2018	R(2+1)D	Sports-1M	ResNet-34	32×10×1	63.6	152×10×1	74.3	91.4
2018	MF-Net	ImageNet	ResNet-34	16×10×5	8	11.1×10×5	72.8	90.4
2019	ip-CSN	Sports-1M	ResNet-101	32×10×3	24.5	83.0×10×3	78.5	93.5
2019	ir-CSN	Sports-1M	ResNet-101	32×10×3	22.1	73.8×10×3	78.1	93.4
2019	SlowFast	-	ResNet-50	(8+32)×10×3	34.4	65.7×10×3	77.0	92.6
2019	SlowFast	-	ResNet-101	(8+64)×10×3	53.7	106×10×3	77.9	93.2
2019	SlowFast	-	ResNet-101+NL	(16+64)×10×3	59.9	234×10×3	79.4	94.4
2020	X3D	-	MobileNet	16×10×3	11	48.4×10×3	79.1	93.9
2021	MoViNet-A6	-	MobileNet	N/A	31.4	386×1×1	81.5	95.3
2018	TSM	ImageNet	ResNet-50	8×10×3	24.3	33×10×3	74.1	-
2018	TSM	ImageNet	ResNet-50	16×10×3	24.3	65×10×3	74.7	91.4
2018	STM	ImageNet	ResNet-50	16×10×3	24	66.5×10×3	73.7	91.6
2018	NL I3D	ImageNet	ResNet-50	128×10×3	35.3	282×10×3	76.5	92.6
2018	NL I3D	ImageNet	ResNet101	128×10×3	54.3	359×10×3	77.7	93.3
2020	TEA	ImageNet	ResNet-50	16×10×3	N/A	70×10×3	76.1	92.5
2020	TDN	ImageNet	ResNet-50	16×10×3	N/A	72×10×3	77.5	93.2
2020	TDN	ImageNet	ResNet-101	16×10×3	N/A	132×10×3	78.5	93.9
2021	VTN	ImageNet	ResNet-50	250×1×1	168	1059×1×1	71.2	90.0
2021	VTN	ImageNet	ResNet-101	250×1×1	187	1989×1×1	72.1	90.3
2021	VTN	ImageNet-21K	ViT-B	250×1×1	114	4218×1×1	78.6	93.7
2021	ViViT-L	JFT	ViT-B	32×3×4	310.8	3992×3×4	81.3	94.7
2021	TokenLearner	JFT	ViT-B	64×3×4	450	4076×3×4	85.4	96.3
2021	MViTv1	-	ViT-B	16×1×5	36.6	70.3×1×5	78.4	93.5
2021	MViTv1	-	ViT-B	32×1×5	36.6	170×1×5	80.2	94.4
2021	MViT-S	-	ViT-B	16×1×5	34.5	64×1×5	81.0	94.6
2021	MViT-B	-	ViT-B	32×1×5	51.2	225×1×5	82.9	95.7
2021	MViT-L	ImageNet-21K	ViT-B	40×3×5	217.6	2828×3×5	86.1	97.0

参考文献：

[1] 马立军. 基于 3D 卷积神经网络的行为识别算法研究 [D]. 北京: 中国地质大学, 2018. (Ma Lijun. Research on action recognition algorithm based on 3d convolutional neural network [D]. Beijing: China University of Geosciences, 2018.)

[2] 何明, 禹明刚, 何红悦, 等. 人工智能的未来 [M]. 北京: 科学出版社, 2020: 31-32. (He Ming, Yu Minggang, He Hongyue, *et al.* The future of artificial intelligence [M]. Beijing: Publishing Sciencep Press, 2020: 31-32.)

[3] Calum Chace. Artificial intelligence and the two singularities [M]. New York: Chapman and Hall/CRC Press, 2018.

[4] 刘勇, 谢若莹, 丰阳, 等. 智能家居中的居民日常行为识别综述 [J]. 计算机工程与应用, 2021, 57 (04): 35-42. (Liu Yong, Xie Ruoying, Feng Yang, *et al.* An overview of resident's daily action recognition in smart home [J]. Computer Engineering and Applications, 2021, 57 (04): 35-42.)

[5] 刘云, 薛盼盼, 李辉, 等. 基于深度学习的关节点行为识别综述 [J]. 电子与信息学报, 2021, 43 (06): 1789-1802. (Liu Yun, Xue Panpan, Li Hui, *et al.* A review of joint behavior recognition based on deep learning [J]. Journal of Electronics & Information Technology, 2021, 43 (06): 1789-1802.)

[6] 张晓平, 纪佳慧, 王力, 等. 基于视频的人体异常行为识别与检测方法综述 [J]. 控制与决策, 2022, 37 (01): 14-27. (Zhang Xiaoping, Ji Jiahui, Wang Li, *et al.* Review of video based human abnormal behavior recognition and detection methods [J]. Control and Decision, 2022, 37 (01): 14-27.)

[7] 裴利沈, 刘少博, 赵雪专. 人体行为识别研究综述 [J]. 计算机科学与探索, 2022, 16 (02): 305-322. (Fei Lishen, Liu Shaobo, Zhao Xuezhuan. A review of human behavior recognition [J]. Journal of Frontiers of Computer Science and Technology, 2022, 16 (02): 305-322.)

[8] Wang H, Schmid C. Action recognition with improved trajectories [C]// Proc of IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2013: 3551-3558.

[9] LeCun Y, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521 (7553): 436-444.

[10] Karpathy A, Toderici G, Shetty S, *et al.* Large-scale video classification with convolutional neural networks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2014: 1725-1732.

[11] Zach C, Pock T, Bischof H. A duality based approach for realtime TV-L1 optical flow [C]// Joint Pattern Recognition Symposium. Berlin: Springer, 2007: 214-223.

[12] He M, Zhu C, Huang Q, *et al.* A review of monocular visual odometry [J]. The Visual Computer, 2020, 36 (5): 1053-1065.

[13] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos [C]. Neural Information Processing Systems. Canada: NIPS Proceedings, 2014, 568-576.

[14] Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016: 1933-1941.

[15] Christoph R, Pinz F A. Spatiotemporal residual networks for video action recognition [C]. Neural Information Processing Systems. Spain: NIPS Proceedings, 2016: 3468-3476.

[16] Feichtenhofer C, Pinz A, Wildes R P. Spatiotemporal multiplier networks for video action recognition [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 7445-7454.

[17] Wang L, Xiong Y, Wang Z, *et al.* Towards Good Practices for Very Deep Two-Stream ConvNets [J]. Computer Science, 2015, 8 (7): 1-5.

[18] Wang L, Qiao Y, Tang X. Action recognition with trajectory-pooled deep-convolutional descriptors [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2015: 4305-4314.

[19] 丁雪琴, 朱轶昇, 朱浩华, 等. 基于时空异构双流卷积网络的行为识别 [J]. 计算机应用与软件, 2022, 39 (03): 154-158. (Ding Xueqin, Zhu Yisheng, Zhu Haohua, *et al.* Action recognition based on spatio-temporal heterogeneous dual-flow convolutional networks [J]. Computer Applications and Software, 2022, 39 (03): 154-158.)

[20] Ji S, Xu W, Yang M, *et al.* 3D convolutional neural networks for human action recognition [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2012, 35 (1): 221-231.

[21] Tran D, Bourdev L, Fergus R, *et al.* Learning spatiotemporal features with 3d convolutional networks [C]// Proc of IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2015: 4489-4497.

[22] Tran D, Ray J, Shou Z, *et al.* Convnet architecture search for spatiotemporal feature learning [J]. Computing Research Repository, 2017, 16 (8): 1-12.

[23] Hara K, Kataoka H, Satoh Y. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018: 6546-6555.

[24] Diba A, Fayyaz M, Sharma V, *et al.* Temporal 3d convnets: New architecture and transfer learning for video classification [EB/OL]. (2017) [2022-04-04]. https://arxiv. org/abs/1711. 08200.

[25] Carreira J, Zisserman A. Quo vadis, action recognition?a new model and the kinetics dataset [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 4724-4733.

[26] Zhou B, Andonian A, Oliva A, *et al.* Temporal relational reasoning in videos [C]// Lecture Notes in Computer Science. Berlin: Springer, 2018: 831-846.

[27] Lin J, Gan C, Han S. Tsm: Temporal shift module for efficient video understanding [C]// Proc of IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2019: 7082-7092.

[28] Shao H, Qian S, Liu Y. Temporal interlacing network [C]// Proc of AAAI Conference on Artificial Intelligence, Palo Alto, CA: AAAI Press, 2020, 34 (07): 11966-11973.

[29] Liu Z, Luo D, Wang Y, *et al.* Teinet: Towards an efficient architecture for video recognition [C]// Proc of AAAI Conference on Artificial Intelligence, Palo Alto, CA: AAAI Press, 2020, 34 (07): 11669-11676.

[30] Liu Z, Wang L, Wu W, *et al.* Tam: Temporal adaptive module for video recognition [C]// Proc of IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2021: 13688-13698.

[31] Li Y, Ji B, Shi X, *et al.* Tea: Temporal excitation and aggregation for action recognition [C]// Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2020: 906-915.

[32] Jiang B, Wang M M, Gan W, *et al.* Stm: Spatiotemporal and motion encoding for action recognition [C]// Proc of IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2019: 2000-2009.

[33] 罗会兰, 陈翰. 时空卷积注意力网络用于动作识别 [J/OL]. 计算机工程与应用, 2022. (2022-03-29) [2022-04-04]. http://kns. cnki. net/kcms/detail/11. 2127. TP. 20220328. 1758. 005. html. (Luo Huilan, Chen Han. Temporal convolution attention network for action

recognition [J/OL]. Computer Engineering and Applications, 2022. (2022-03-29) [2022-04-04]. http://kns. cnki. net/kcms/detail/11. 2127. TP. 20220328. 1758. 005. html.)

[34] 吴丽君，李斌斌，陈志聪，等. 3D 多重注意力机制下的行为识别 [J]. 福州大学学报（自然科学版）, 2022, 50 (01): 47-53. (Wu Lijun, LI Binbin, Chen Zhicong, *et al.* Action recognition based on 3D multi-attention mechanism [J]. Journal of Fuzhou University (Natural Science) , 2022, 50 (01): 47-53.)

[35] Wang L, Xiong Y, Wang Z, *et al.* Temporal segment networks: Towards good practices for deep action recognition [C]// Lecture Notes in Computer Science. Berlin: Springer, 2016: 20-36.

[36] Zhu W, Hu J, Sun G, *et al.* A key volume mining deep framework for action recognition [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016: 1991-1999.

[37] Liu X, Pintea S L, Nejadasl F K, *et al.* No frame left behind: Full Video Action Recognition [C]// Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2021: 14887-14896.

[38] Gowda S N, Rohrbach M, Sevilla-Lara L. SMART Frame Selection for Action Recognition [C]// Proc of AAAI Conference on Artificial Intelligence, Palo Alto, CA: AAAI Press, 2021, 35 (2): 1451-1459.

[39] Dosovitskiy A, Fischer P, Ilg E, *et al.* Flownet: Learning optical flow with convolutional networks [C]// Proc of IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2015: 2758-2766.

[40] Ilg E, Mayer N, Saikia T, *et al.* Flownet 2. 0: Evolution of optical flow estimation with deep networks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 1647-1655.

[41] Piergiovanni A J, Ryoo M S. Representation flow for action recognition [C]// Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2019: 9937-9945.

[42] Zhu Y, Lan Z, Newsam S, *et al.* Hidden two-stream convolutional networks for action recognition [C]// Lecture Notes in Computer Science. Berlin: Springer, 2018: 363-378.

[43] Crasto N, Weinzaepfel P, Alahari K, *et al.* Mars: Motion-augmented rgb stream for action recognition [C]// Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2019: 7874-7883.

[44] Zhang C, Zou Y, Chen G, *et al.* Pan: Towards fast action recognition via learning persistence of appearance [EB/OL]. (2020) [2022-04-04]. https://arxiv. org/abs/2008. 03462.

[45] Qiu Z, Yao T, Mei T. Learning spatio-temporal representation with pseudo-3d residual networks [C]// Proc of IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2017: 5534-5542.

[46] Tran D, Wang H, Torresani L, *et al.* A closer look at spatiotemporal convolutions for action recognition [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018: 6450-6459.

[47] Xie S, Sun C, Huang J, *et al.* Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification [C]// Lecture Notes in Computer Science. Berlin: Springer, 2018: 318-335.

[48] Sudhakaran S, Escalera S, Lanz O. Gate-shift networks for video action recognition [C]// Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2020: 1099-1108.

[49] Chen Y, Kalantidis Y, Li J, *et al.* Multi-fiber networks for video recognition [C]// Lecture Notes in Computer Science. Berlin: Springer, 2018: 364-380.

[50] Tran D, Wang H, Torresani L, *et al.* Video classification with channel-separated convolutional networks [C]// Proc of IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2019: 5551-5560.

[51] Luo C, Yuille A L. Grouped spatial-temporal aggregation for efficient action recognition [C]// Proc of IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2019: 5511-5520.

[52] Zhou Y, Sun X, Zha Z J, *et al.* Mict: Mixed 3d/2d convolutional tube for human action recognition [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018: 449-458.

[53] Zolfaghari M, Singh K, Brox T. Eco: Efficient convolutional network for online video understanding [C]// Lecture Notes in Computer Science. Berlin: Springer, 2018: 713-730.

[54] Wang L, Li W, Li W, *et al.* Appearance-and-relation networks for video classification [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018: 1430-1439.

[55] Feichtenhofer C, Fan H, Malik J, *et al.* Slowfast networks for video recognition [C]// Proc of IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2019: 6201-6210.

[56] Yang C, Xu Y, Shi J, *et al.* Temporal pyramid network for action recognition [C]// Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2020: 588-597.

[57] Huang G, Bors A G. Busy-Quiet Video Disentangling for Video Classification [C]// Proc of IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway, NJ: IEEE Press, 2022: 756-765.

[58] 刘钊，杨帆，司亚中. 时域非填充网络视频行为识别算法研究 [J/OL]. 计算机工程与应用, 2022. (2022-01-16) [2022-04-04]. http://kns. cnki. net/kcms/detail/11. 2127. TP. 20220106. 1220. 002. html. (Liu Zhao, Yang Fan, Si Yazhong. Research on time-domain unfilled network video behavior recognition Algorithm [J]. Computer Engineering and Applications, 2022. (2022-01-16) [2022-04-04]. http://kns. cnki. net/kcms/detail/11. 2127. TP. 20220106. 1220. 002. html.)

[59] Feichtenhofer C. X3d: Expanding architectures for efficient video recognition [C]// Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2020: 200-210.

[60] Kondratyuk D, Yuan L, Li Y, *et al.* Movinets: Mobile video networks for efficient video recognition [C]// Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2021: 16015-16025.

[61] Varol G, Laptev I, Schmid C. Long-term temporal convolutions for action recognition [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2017, 40 (6): 1510-1517.

[62] Lan Z, Zhu Y, Hauptmann A G, *et al.* Deep local video feature for action recognition [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE Press, 2017: 1219-1225.

[63] Girdhar R, Ramanan D, Gupta A, *et al.* Actionvlad: Learning spatio-temporal aggregation for action classification [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 3165-3174.

[64] Diba A, Sharma V, Van Gool L. Deep temporal linear encoding networks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 1541-1550.

[65] Wang L, Tong Z, Ji B, *et al.* TDN: Temporal difference networks for efficient action recognition [C]// Proc of the IEEE/CVF Conference on

Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2021: 1895-1904.

[66] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural computation, 1997, 9 (8): 1735-1780.

[67] Srivastava N, Mansimov E, Salakhudinov R. Unsupervised learning of video representations using lstms [C]// Proc of the 32th International Conference on Machine Learning. Cambridge MA: JMLR, 2015: 843-852.

[68] Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, *et al.* Beyond short snippets: Deep networks for video classification [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2015: 4694-4702.

[69] Donahue J, Anne Hendricks L, Guadarrama S, *et al.* Long-term recurrent convolutional networks for visual recognition and description [J]// IEEE Trans on Pattern Analysis and Machine Intelligence, 2017, 39 (4): 677-691.

[70] Ma C Y, Chen M H, Kira Z, *et al.* TS-LSTM and temporal-inception: exploiting spatiotemporal dynamics for activity recognition [J]. Signal Processing: Image Communication, 2019, 71: 76-87.

[71] Wang X, Miao Z, Zhang R, *et al.* I3D-LSTM: A new model for human action recognition [C/OL]// Proc of IOP Conference Series: Materials Science and Engineering. S. l. : IOP, 2019, 569 (3): 032035.

[72] Li Z, Gavrilyuk K, Gavves E, *et al.* Videolstm convolves, attends and flows for action recognition [J]. Computer Vision and Image Understanding, 2018, 166: 41-50.

[73] Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need [C]// Neural Information Processing Systems. USA: NIPS Proceedings, 2017: 5998-6008.

[74] Wang X, Girshick R, Gupta A, *et al.* Non-local neural networks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018: 7794-7803.

[75] Neimark D, Bar O, Zohar M, *et al.* Video transformer network [C]// Proc of IEEE/CVF International Conference on Computer Vision Workshops. Piscataway, NJ: IEEE Press, 2021: 3156-3165.

[76] Li K, Wang Y, Gao P, *et al.* Uniformer: Unified Transformer for Efficient Spatiotemporal Representation Learning [EB/OL]. (2022) [2022-04-04]. https://arxiv. org/abs/2201. 04676.

[77] Arnab A, Dehghani M, Heigold G, *et al.* Vivit: A video vision transformer [C]// Proc of IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2021: 6816-6826.

[78] Dosovitskiy A, Beyer L, Kolesnikov A, *et al.* An image is worth 16x16 words: Transformers for image recognition at scale [EB/OL]. (2020) [2022-04-04]. https://arxiv. org/abs/2010. 11929.

[79] Fan H, Xiong B, Mangalam K, *et al.* Multiscale vision transformers [C]// Proc of IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2021: 6804-6815.

[80] Li Y, Wu C Y, Fan H, *et al.* Improved multiscale vision transformers for classification and detection [EB/OL]. (2021) [2022-04-04]. https://arxiv. org/abs/2112. 01526.

[81] Zhang Y, Li X, Liu C, *et al.* Vidtr: Video transformer without convolutions [C]// Proc of IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2021: 13557-13567.

[82] Sharir G, Noy A, Zelnik-Manor L. An Image is Worth 16x16 Words, What is a Video Worth? [EB/OL]. (2021) [2022-04-04]. https://arxiv. org/abs/2103. 13915.

[83] Truong T D, Bui Q H, Duong C N, *et al.* DirecFormer: A Directed Attention in Transformer Approach to Robust Action Recognition [EB/OL]. (2022) [2022-04-04]. https://arxiv. org/abs/2203. 10233

[84] Wang R, Chen D, Wu Z, *et al.* Bevt: Bert pretraining of video transformers [EB/OL]. (2021) [2022-04-04]. https://arxiv. org/abs/2112. 01529.

[85] Alfasly S, Lu J, Xu C, *et al.* Learnable Irrelevant Modality Dropout for Multimodal Action Recognition on Modality-Specific Annotated Videos [EB/OL]. (2022) [2022-04-04]. https://arxiv. org/abs/2203. 03014.

[86] Zellers R, Lu J, Lu X, *et al.* MERLOT Reserve: Neural Script Knowledge through Vision and Language and Sound [EB/OL]. (2022) [2022-04-04]. https://arxiv. org/abs/2201. 02639.

[87] Duan H, Zhao Y, Xiong Y, *et al.* Omni-sourced webly-supervised learning for video recognition [C]// Lecture Notes in Computer Science. Berlin: Springer, 2020: 670-688.

[88] Zhang B, Yu J, Fifty C, *et al.* Co-training Transformer with Videos and Images Improves Action Recognition [EB/OL]. (2021) [2022-04-04]. https://arxiv. org/abs/2112. 07175.

[89] Ryoo M S, Piergiovanni A J, Arnab A, *et al.* TokenLearner: What Can 8 Learned Tokens Do for Images and Videos? [EB/OL]. (2021) [2022-04-04]. https://arxiv. org/abs/2106. 11297.

[90] Jhuang H, Garrote H, Poggio E, *et al.* HMDB: A large video database for human motion recognition [C]// Proc of IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2011: 2556-2563.

[91] Soomro K, Zamir A R, Shah M. A dataset of 101 human action classes from videos in the wild [J]. Center for Research in Computer Vision, 2012, 2 (11) .

[92] Kay W, Carreira J, Simonyan K, *et al.* The kinetics human action video dataset [EB/OL]. (2017) [2022-04-04]. https://arxiv. org/abs/1705. 06950

[93] Carreira J, Noland E, Banki-Horvath A, *et al.* A short note about kinetics-600 [EB/OL]. (2018) [2022-04-04]. https://arxiv. org/abs/1808. 01340.

[94] Carreira J, Noland E, Hillier C, *et al.* A short note on the kinetics-700 human action dataset [EB/OL]. (2019) [2022-04-04]. https://arxiv. org/abs/1907. 06987.

[95] Smaira L, Carreira J, Noland E, *et al.* A short note on the kinetics-700-2020 human action dataset [EB/OL]. (2020) [2022-04-04]. https://arxiv. org/abs/2010. 10864.

[96] Goyal R, Ebrahimi Kahou S, Michalski V, *et al.* The"something something"video database for learning and evaluating visual common sense [C]// Proc of IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2017: 5843-5851.

[97] Marszalek M, Laptev I, Schmid C. Actions in context [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2009: 2929-2936.

[98] Karpathy A, Toderici G, Shetty S, *et al.* Large-scale video classification with convolutional neural networks [C]// Proc of the Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2014: 1725-1732.

[99] Caba Heilbron F, Escorcia V, Ghanem B, *et al.* Activitynet: A large-scale video benchmark for human activity understanding [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2015: 961-970.

[100] [100] Sigurdsson G A, Varol G, Wang X, *et al.* Hollywood in homes: Crowdsourcing data collection for activity understanding [C]// Lecture Notes in Computer Science. Berlin: Springer, 2016: 510-526.

[101] [101] Monfort M, Andonian A, Zhou B, *et al.* Moments in time dataset: one million videos for event understanding [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2019, 42 (2): 502-508.

[102] [102] Kong Q, Wu Z, Deng Z, *et al.* Mmact: A large-scale dataset for cross modal human action understanding [C]// Proc of IEEE/CVF

International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2019: 8657-8666.

[103] [103] Materzynska J, Berger G, Bax I, *et al.* The jester dataset: A large-scale video dataset of human gestures [C]// Proc of IEEE/CVF International Conference on Computer Vision Workshops. Piscataway, NJ: IEEE Press, 2019: 2874-2882.

[104] [104] Malla S, Dariush B, Choi C. Titan: Future forecast using action priors [C]// Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2020: 11183-11193.

[105] [105] Miech A, Alayrac J B, Laptev I, *et al.* RareAct: A video dataset of unusual interactions [EB/OL]. (2020) [2022-04-04]. https://arxiv. org/abs/2008. 01018.

[106] [106] Demir U, Rawat Y S, Shah M. TinyVIRAT: low-resolution video action recognition [C]// Proc of the 25th International Conference on Pattern Recognition. Piscataway, NJ: IEEE Press, 2021: 7387-7394.

[107] [107] Li T, Liu J, Zhang W, *et al.* UAV-Human: A Large Benchmark for Human Behavior Understanding with Unmanned Aerial Vehicles [C]// Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2021: 16261-16270.

[108] [108] Rai N, Chen H, Ji J, *et al.* Home action genome: Cooperative compositional action understanding [C]// Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2021: 11179-11188.

[109] [109] Thatipelli A, Narayan S, Khan S, *et al.* Spatio-temporal Relation Modeling for Few-shot Action Recognition [EB/OL]. (2021) [2022-04-04]. https://arxiv. org/abs/2112. 05132.

[110] [110] Huang Z, Zhang S, Jiang J, *et al.* Self-supervised motion learning from static images [C]// Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2021: 1276-1285.

[111] [111] Wei C, Fan H, Xie S, *et al.* Masked Feature Prediction for Self-Supervised Visual Pre-Training [EB/OL]. (2021) [2022-04-04]. https://arxiv. org/abs/2112. 09133.